An Automatic Traffic Congestion Identification Algorithm Based on Mixture of Linear Regressions

Mohammed Elhenawy, Hesham Rakha(✉), and Hao Chen

Virginia Tech Transportation Institute,
3500 Transportation Research Plaza, Blacksburg, VA 24061, USA
{elhenawy, hrakha, hchen}@vt.edu

Abstract. One innovative solution to traffic congestion is to use real-time data and Intelligent Transportation Systems (ITSs) to optimize the existing transportation system. To address this need, we propose an algorithm for real-time automatic congestion identification that uses speed probe data and the corresponding weather and visibility to build a unified model. Based on traffic flow theory, the algorithm assumes three traffic states: congestion, speed-at-capacity, and free-flow. Our algorithm assumes that speed is drawn from a mixture of three components, whose means are functions of weather and visibility and defined using a linear regression of their predictors. The parameters of the model were estimated using three empirical datasets from Virginia, California, and Texas. The fitted model was used to calculate the speed cut-off between congestion and speed-at-capacity by minimizing either the Bayesian classification error or the false positive (congestion) rate. The test results showed promising congestion identification performance.

Keywords: Transportation planning and traffic operation · Real-time automatic congestion identification · Mixture of linear regression · ITS

1 Introduction

Traffic congestion has become one of the problems of modern life in many metropolitan areas. This growing problem has environmental effects. During congestion, cars cannot run efficiently, so air pollution, carbon dioxide (CO2) emissions, and fuel use increase. In 2007, Americans lost $87.2 billion in wasted fuel and lost productivity from congestion. This waste reached $115 billion in 2009 [1]. Congestion also increases travel time; for example, in 1993 driving under congested conditions caused a delay of about 0.6 min per km of travel on expressways and 1.2 min of delay per km of travel in arterials [2]. The congestion problem has only grown since then. The Texas Transportation Institute has reported that the number of hours Americans wasted in traffic congestion increased fivefold between 1982 and 2005. Moreover, congestion has an economic effect. Studies show that congestion slows metropolitan growth, inhibits agglomeration economies, and shapes economic geographies [3]. Traffic congestion can be caused by an obstruction or lack of road capacity, which is an inefficient use of the roads. This problem can be reduced by increasing budgets for the

M. Helfert et al. (Eds.): SMARTGREENS 2016 and VEHITS 2016, CCIS 738, pp. 242–256, 2017.
DOI: 10.1007/978-3-319-63712-9_14

construction of roads and infrastructure. But adding more road capacity is costly, budgets are limited, and the construction itself takes a long time.

With the continuous increase in traffic volumes, managing traffic, particularly at times of peak demand, is a good and inexpensive solution to congestion. Advanced traffic management systems (ATMS) use various applications of intelligent transportation systems (ITS) to manage traffic and reduce congestion problems. Recently, advancements in communication and computers have greatly improved ITS and made it more capable of identifying and reducing congestion. ITS is an effective solution to traffic problems because it improves the dynamic capacity of the road system without building extra expensive infrastructure [4]. Accurate and real-time traffic information is the foundation of ITS.

Congestion usually starts from a road bottleneck, then spills over to neighboring road segments. It takes time until this congestion disappears. Depending on the frequency of occurrence, traffic congestion can be divided into two categories [5]. The first is recurrent traffic congestion, and the second is accidental (non-recurring) traffic congestion. Recurrent traffic congestion, which usually results from exceeding the road capacity, is easier to identify and predict. Accidental traffic congestion usually results from traffic incidents or severe weather conditions. Traffic congestion is different for different locations, time periods, and weather conditions.

The impact of weather on freeway traffic operations is a major concern of roadway management agencies; however, little research has been done to link weather and congestion in a quantitative sense. Two groups at the University of Washington correlated weather and traffic phenomena using the Traffic Data Acquisition and Distribution (TDAD) data mine and the Doppler radar data mine [6]. Their basic idea was that if moving weather cells could be tracked and predicted using weather radar then a correlation between the properties of the weather cell and observed traffic states could be found. Nookala studied the traffic congestion caused by weather conditions and their effect on traffic volume and travel time [7]. He observed an increase in the traffic congestion during inclement weather conditions resulting from a drop in freeway capacity without a corresponding significant drop in traffic demand. Chung et al. used traffic data collected over a 2-year period from July 1, 2002, to June 30, 2004, at the Tokyo Metropolitan Expressway (MEX) and showed a decrease in free-flow speed and in capacity with increasing amount of rainfall [8]. Brilon and Ponzlet used 3 years of historical data for 15 freeway sites in Germany to investigate the impacts of several factors, including weather, on speed-flow relationships [9]. They found that wet roadway conditions cause different reductions in speed on highways with different numbers of lanes. Agarwal et al. emphasized that the results obtained from studies outside the United States cannot be applied within the United States due to different roadway and driver characteristics. Moreover, the results obtained from rural freeway segments within the United States may be different from urban freeways [10]. Ibrahim and Hall used a limited historical dataset and multiple regression analysis to study the impact of rain and snow on speed [1]. Their results showed that light rain and snow cause similar reductions in speeds (3%–5%), while 14%–15% and 30%–40% reductions in speed are caused by heavy rain and heavy snow respectively. Rakha et al. used weather data (precipitation and visibility) and loop detector data (speed, flow, and occupancy) obtained from Baltimore, Maryland, Minneapolis/St. Paul, Minnesota, and Seattle, Washington, in the United States to quantify the impact of inclement weather on traffic stream behavior and key traffic stream parameters, including free-flow speed, speed-at-capacity, capacity, and jam density [11].

During the last few years, many automatic congestion identification algorithms have been proposed. ASBIA is an algorithm that uses speed measurements over short temporal intervals and spatial segments to identify the status of a segment using the *t*-test [12]. The outputs of the algorithm are the status of the roadway segment (free-flow or congested) and the confidence level of the test (*p*-value). Another algorithm uses vehicle trajectories in an intelligent vehicle infrastructure co-operation system (IVICS) [4]. The spatial-temporal trajectories are considered as an image to extract the propagation speed of a congestion wave and construct a congestion template. The correlation is evaluated between the template and the spatial-temporal velocity image to identify the congestion. A parallel support vector machine (SVM) is used in [13] to identify traffic congestion. The authors proposed parallel SVM instead of SVM because the training computation cost of SVM is expensive and congestion identification is a real-time task.

Floating car data are used in [14] to find meaningful congestion patterns. The analysis of the floating car data is done using a method based on a data cube and the spatial-temporal related relationship of the slow-speed road segment to identify the traffic congestion. The research team at the Center for Sustainable Mobility (CSM) at the Virginia Tech Transportation Institute (VTTI) developed an algorithm to identify congested segments using a spatiotemporal speed matrix [15]. The proposed algorithm fits two log-normal (or normal) distributions to the training dataset.

To the best of our knowledge, no research addresses the impacts of both visibility and weather conditions on congestion identification. In this paper, the impacts of weather conditions and visibility levels on the congestion identification algorithm are investigated by modeling the speed distribution as a mixture of three log-normal components whose means are linear functions of weather condition and visibility level. So that based on weather condition and visibility level the three log-normal components may get close or apart and the cut-off speed is changed. The proposed algorithm was built using three different datasets from three different states (Virginia, Texas, and California). The results of our proposed model are promising and reasonable; for example, the cut-off speed increases as the visibility level increases.

The remainder of this paper is organized as follows. First, a brief background of the method used is given. After that, the proposed algorithm is introduced. The datasets used in the case study are described. Subsequently, the result of the experimental work is explained and an illustrative example is given to show how to implement the proposed model. Finally, conclusions are presented.

2 Mixture of Linear Regressions

Finite mixture models are powerful tools for analyzing a wide variety of random phenomena. They are used to model random phenomena in many fields, including agriculture, biology, economics, medicine, and genetics. A mixture of linear regressions is one of the mixture families that has been studied carefully in the literature [16, 17]. It can be used to model the speed for different traffic regimes at different weather condition and visibility levels.

The mixture of linear regression can be written as:

$$p(y|X) = \sum_{j=1}^{m} \frac{\lambda_j}{\sigma_j\sqrt{2\pi}} e^{-\frac{\left(y - X^T\beta_j\right)^2}{2\sigma_j^2}} \quad (1)$$

Or as

$$y_i = \begin{cases} x_i^T\beta_1 + \in_{i1} \text{ with probability } \lambda_1 \\ x_i^T\beta_2 + \in_{i2} \text{ with probability } \lambda_2 \\ \cdot \\ \cdot \\ \cdot \\ x_i^T\beta_m + \in_{im} \text{ with probability } 1 - \sum_{q=1}^{m-1}\lambda_q \end{cases} \quad (2)$$

where y_i is a response corresponding to a predictors vector x_i^T, β_j is a vector of regression coefficients for the j^{th} mixture component, λ_j is a mixing probability of the j^{th} mixture component, $\in_{ij}$ are normal random errors, and m is the number of components in the mixture model. Model parameters $\psi = \left\{\beta_1, \beta_2, \ldots, \beta_m, \sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2, \lambda_1, \lambda_2, \ldots, \lambda_m\right\}$ can be estimated by maximizing the log-likelihood of Eq. (1) given a set of response predictor pairs, $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$, and using the Expectation-Maximization algorithm (EM).

2.1 EM Algorithm

The EM algorithm iteratively finds maximum likelihood estimates by alternating the E-step and M-step. Let $\psi^{(k)}$ be parameter estimates after the k^{th} iteration. On the E-step, the posterior probability of the i^{th} observation comes from component j and is computed as shown in Eq. (3).

$$w_{ij}^{(k+1)} = \frac{\lambda_j^{(k)}\phi_j\left(y_i|x_i, \psi^{(k)}\right)}{\sum_{j=1}^{m}\lambda_j^{(k)}\phi_j\left(y_i|x_i, \psi^{(k)}\right)} \quad (3)$$

where $\phi_j\left(y_i|x_i, \psi^{(k)}\right)$ is the probability density function of the j^{th} component.

On the M-step, new parameter estimates $\psi^{(k+1)}$ maximizing the log-likelihood function in Eq. (1) are calculated, as shown in Eqs. (4) and (5).

$$\lambda_j^{(k+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(k+1)}}{n} \quad (4)$$

$$\widehat{\beta}_j^{(k+1)} = (X^T W_j X)^{-1} X^T W_j Y \quad (5)$$

where $X_{nx(p+1)}$ is the predictors matrix, Y_{nx1} is the corresponding response vector, and W_{nxn} is an diagonal matrix having $w_{ij}^{(k+1)}$ along its diagonal.

$$\hat{\sigma}_j^{2(k+1)} = \frac{\sum_{i=1}^{n} w_{ij}^{(k+1)} (y_i - x_i^T \hat{\beta}_j^{(k+1)})^2}{\sum_{i=1}^{n} w_{ij}^{(k+1)}} \tag{6}$$

The E-step and M-step are alternated repeatedly until the incomplete log-likelihood change is arbitrarily small, as shown in Eq. (7).

$$\left| \prod_{i=1}^{n} \sum_{j=1}^{m} \lambda_j^{(k+1)} \phi_j \left(y_i | x_i, \psi^{(k+1)} \right) - \prod_{i=1}^{n} \sum_{j=1}^{m} \lambda_j^{(k)} \phi_j \left(y_i | x_i, \psi^{(k)} \right) \right| < \xi \tag{7}$$

where ξ is a small number.

3 The Proposed Algorithm

The proposed algorithm is the result of a research effort that extended over a couple of years. Our research efforts resulted in the development of three algorithms. The first algorithm is based on the one-sample *t*-test [12]. This algorithm uses speed measurements over short temporal intervals and spatial segments to identify the status of the center point as being in a free-flow or congested state using a *t*-test. The drawbacks of this algorithm are its need of future speed readings, a normality assumption, and multiple testing corrections. In order to overcome the drawbacks of this algorithm, we proposed a second algorithm.

The second algorithm fits a mixture of two components using a historical dataset [15]. The fitted mixture model is used to calculate a threshold to distinguish between free-flow and congested traffic. If the speed of a segment is greater than this threshold, the segment is considered to be in a free-flow state; otherwise the segment is considered congested. The second algorithm overcomes the model deficiencies of the *t*-test-based algorithm, and it overcomes the normality problem by using the log-normal distribution to model skewed data. It is also suitable for online (real-time) application because it does not require knowledge of future speed readings. However, the drawback of this algorithm is that it does not consider any weather conditions or visibility levels.

The third algorithm uses a mixture of two linear regressions to model the speed distributions for different traffic conditions, including free-flow and congested traffic [18]. The speed of each regime is modeled as a distribution whose mean is function of weather condition and visibility level. The threshold that separates free-flow and congested traffic becomes a function of weather and visibility level as well. The proposed model overcomes the problem of limited data for some weather conditions and visibility levels by pooling the data during model fitting. However, this algorithm has a serious problem in that it overestimates the thresholds separating the free-flow and congested regimes. To overcome this problem, we used the traffic flow theory fundamental diagram, which assumes that there are three traffic states: free-flow, speed-at-capacity, and congested.

As shown in the fundamental diagrams in Fig. 1, we divide the traffic states of a road segment into three traffic regimes where the speed of each regime can be modeled by a log-normal distribution. Consequently, the overall speed distribution can be represented as a mixture of three log-normal components. The first regime is free-flow, which has the speed distribution with the highest mean. At free-flow, traffic density lies below the capacity density. The second regime is congested flow, which has the speed distribution with the lowest mean. Congested flow is characterized by traffic density between the capacity density and the jam density. The third regime is capacity flow, which separates free-flow from congested flow. Its speed distribution has a mean between the means of the other two regimes.

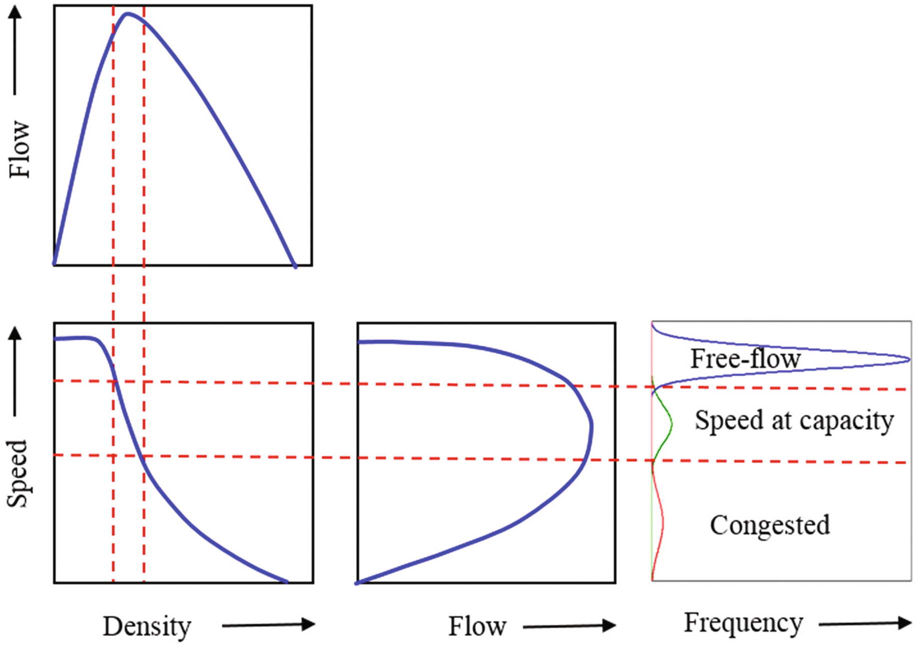


Fig. 1. Illustration of link between the fundamental diagrams and the three components mixture.

As has been shown in several studies, the flow fundamental diagram is affected by weather conditions [11, 19, 20]. Thus, we expect the mean of the speed distribution corresponding to each regime to change with weather and visibility. The proposed algorithm uses a mixture of three linear regressions and real datasets to learn the means of the distribution as a function of weather and visibility and find the boundary between the three regimes. The proposed algorithm is shown below in Table 1.

All segments with speeds greater than the threshold are classified as free-flow segments, and other segments are classified as congested segments. The output of the above algorithm is a spatiotemporal binary matrix with dimensions identical to the spatiotemporal speed matrix. A "1" in the binary matrix identifies a segment as congested, and a "0" represents free-flow conditions.

Table 1. The proposed algorithm.

1. Use the EM algorithm described earlier to fit three component distributions to locally collected data, as demonstrated in Equation (8).

$$(\log(y) \mid \lambda_1, \lambda_2, \beta_1, \beta_2, \beta_3, \sigma_1, \sigma_2, \sigma_3) = \lambda_1 \frac{1}{\sqrt{2\pi}\sigma_1} e^{\frac{\left(\log(y) - X^T\beta_1\right)^2}{2\sigma_1^2}} + \lambda_2 \frac{1}{\sqrt{2\pi}\sigma_2} e^{\frac{\left(\log(y) - X^T\beta_2\right)^2}{2\sigma_2^2}} + (1 - \lambda_2 - \lambda_1) \frac{1}{\sqrt{2\pi}\sigma_3} e^{\frac{\left(\log(y) - X^T\beta_3\right)^2}{2\sigma_3^2}} \quad (1)$$

   where vector *X* is a vector of weather conditions and visibility predictors.
   Here $(X^T\beta_1, \sigma_1)$, $(X^T\beta_2, \sigma_2)$, and $(X^T\beta_3, \sigma_3)$ are the locations and spreads of the mixture components, and (λ_1, λ_2) are the mixture parameters.
2. For unseen data, use the weather condition, visibility level, and equations of the means ($X^T\beta_1$, $X^T\beta_2$, and $X^T\beta_3$) to calculate locations (means) of the three components.
3. Calculate the cut-off speed. We have two options to calculate the cut-off speed and we can use either of them:
   (a) Calculate the 0.001 quintile of the speed-at-capacity (the middle distribution).
   (b) Calculate cut-off speed using the Bayesian approach, which finds the intersection point (between congestion and speed-at-capacity) that minimizes classification error [18].
4. Use cut-off speed as a threshold to classify the state of each road segment.

4 Experimental Work

4.1 Data Reduction

In order to use collected traffic data in the proposed algorithm, data reduction was an important process for transforming the raw measured data into the required data input formats. In general, the spatiotemporal traffic state matrix is a fundamental attribute of the input data. Reduction of INRIX probe data is one example, and a similar process can be applied to other types of measured data (e.g., loop detector). INRIX data are collected for each roadway segment and time interval. Each roadway segment represents a traffic management center (TMC) station. Geographic TMC station information is also provided. The average speed for each TMC station can be used to derive a spatiotemporal traffic state matrix. However, raw INRIX data includes geographically inconsistent sections, irregular data collection time intervals, and missing data. Considering these problems, the data reduction process is illustrated in Fig. 2.

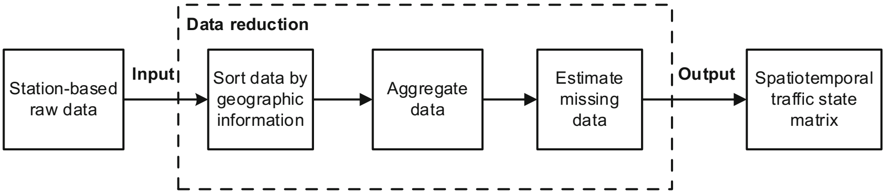


Fig. 2. Data reduction of INRIX probe data.

Based on the geographic information of each TMC station, raw data are sorted by roadway direction (e.g., eastbound or westbound). As part of this step, the data should be checked for any overlapping or inconsistent stations along the direction. Afterwards, speed data should be aggregated by time intervals (e.g., 5 min), according to the algorithm's resolution requirement. In this way, raw data can be aggregated into a daily matrix format along spatial and temporal intervals. It should be noted that missing data usually exist in the developed data matrix. Therefore, data imputation methods should be conducted to estimate the missing data based on neighboring cell values. Consequently, the daily spatiotemporal traffic state matrix can be generated for congestion and bottleneck identification.

4.2 Study Sites

INRIX traffic data from three states (Virginia, Texas, and California) were used to develop the proposed automatic congestion identification algorithm. Specifically, the study included 2011 ∼ 2013 data along I-66 eastbound, 2012 data along US-75 northbound, and 2012 data along I-15 southbound. The selected freeway corridor on I-66 is presented in Fig. 3, which includes 36 freeway segments along 30.7 miles. Average speeds (or travel times) for each roadway segment are provided in the raw data, which were collected every minute. In order to reduce the stochastic noise and measurement error, raw speed data were aggregated by 5-minute intervals. Therefore, the traffic speed matrix over spatial (upstream to downstream) and temporal (from 0:00 to 23:55) domains could be obtained for each day. For the other two locations, daily speed matrices were obtained using the same procedure. Selected freeway corridors on US-75 and I-15, which include 81 segments across 38 miles and 30 segments across 15.6 miles, respectively, are presented in Figs. 4 and 5.

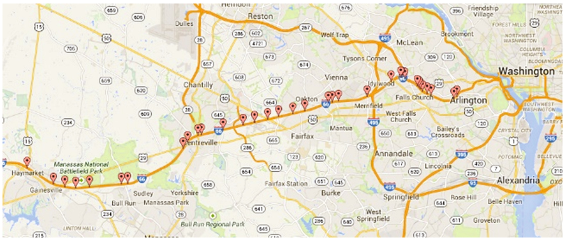


Fig. 3. Layout of the selected freeway stretch on I-66 (Source: Google Maps).

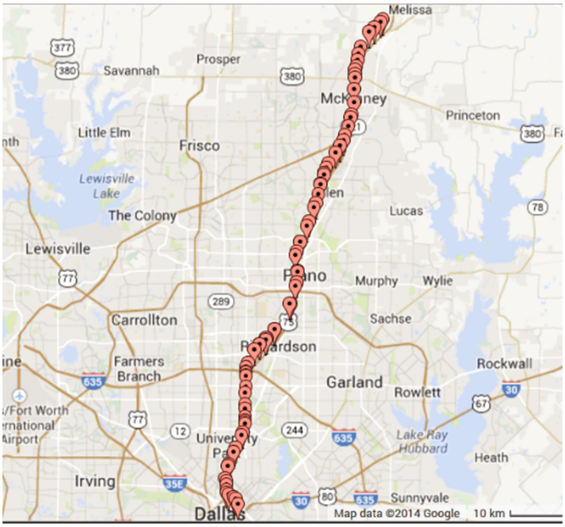


Fig. 4. Layout of the selected freeway stretch on US-75 (Source: Google Maps).

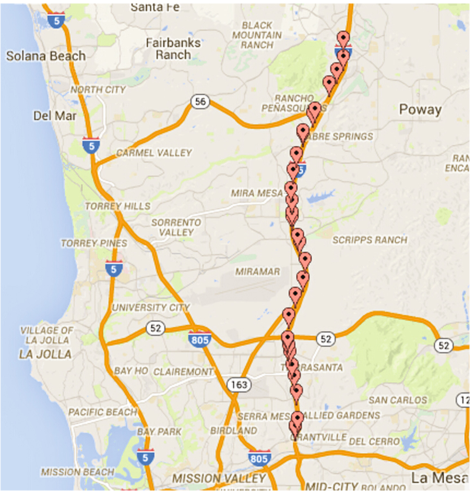


Fig. 5. Layout of the selected freeway stretch on I-15 (Source: Google Maps).

4.3 Effect of Visibility and Weather Conditions

This subsection describes the investigation of weather and visibility impacts on the cut-off speed (threshold) that is used to define the congested condition. The investigation was limited by the fact that data could not be divided into bins containing each weather condition and visibility level. Moreover, many bins had small amounts of data or no data at all. With this in mind, a mixture of linear regressions was used to pool data and estimate cut-off speeds, without sorting the data into clusters. In this subsection, we describe a speed model featuring a mix of three linear regressions. Each linear equation describes a relationship between independent variables (visibility and weather) and the dependent variable, which is speed. In other words, instead of mixing three components with unchanged means, the speed model mixes three components whose means are a function of weather and visibility.

4.4 Unified Model

In order to get a unified model that is independent of the location or the speed limit, we did the following:

1. Weather conditions for the three datasets were consolidated based on precipitation, as shown in Appendix A. Weather conditions from all three datasets were then mapped onto these weather groups.
2. We put all three datasets in one pool and did not include indicator variables that would identify the dataset.
3. The speed was normalized by dividing the speed at each road segment by the posted maximum speed at this segment.

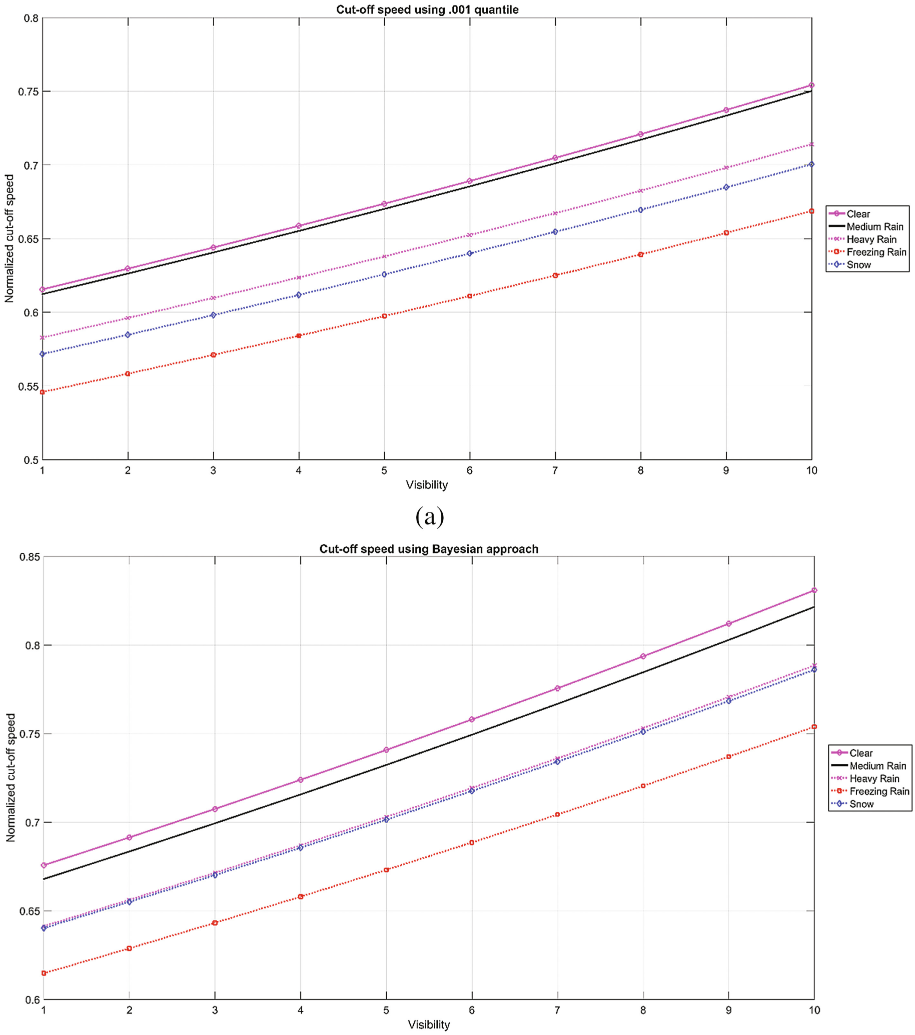
The unified model has a response which is the normalized speed coming from the three datasets. The predictors are the indicator variables for the weather groups and the visibility level.

Before applying the mixture of three linear regression models was applied, speed and visibility data were grouped by weather. Because the dataset was huge and we could not estimate the model parameters using the whole dataset at once due to memory issues, a total of 7,000 random samples were drawn randomly from each weather group to construct a realization (dataset). Each random sample included the speed and visibility level, together with indicator variables for the weather. Because speed distributions are skewed, the log-normal distribution is preferred to the normal distribution. Log speed was used as the response variable. Weather code and visibility were the explanatory variables (predictors). Coefficients of the predictors $(\beta_1, \beta_2, \beta_3)$, the variance of each component $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$, and the proportions $(\lambda_1, \lambda_2, \lambda_3)$ of each component were estimated using the iterative EM algorithm (Eqs. 3–6). This procedure was repeated 300 times by bootstrapping the sample construction without replacement. Final model parameters were the mean or median of all model coefficients. Once the final model was derived, we could observe the shift of the distribution mean with the weather condition and visibility level in the three regimes (free-flow, speed-at-capacity, and congested). Given any combination of weather and visibility, the final model computes mean speeds for the three regimes. Furthermore, using the estimated model's parameters, the model computes Bayesian and 0.001 quantile cut-off speeds.

The estimated general model's parameters are shown in Table 2. As shown in Fig. 6, the results are sensible because all weather groups have cut-off speeds lower than or equal to the clear group. Moreover, the cut-off speed increases as visibility increases. We should mention that the cut-off speeds for clear and light rain are very close, so we can apply the cut-off speed of the clear condition to light rain as well. Appendix B shows the speed matrix and the corresponding binary matrix after applying the proposed algorithm.

Table 2. Unified model's parameters.

	Congestion	Speed-at-capacity	Free-flow
"Clear" (Intercept)	−0.9025	−0.1947	0.0335
"Visibility"	0.0260	0.0229	0.0026
"Medium Rain"	−0.0722	−0.0024	−0.0238
"Heavy Rain"	−0.0398	−0.0465	−0.0308
"Freezing Rain"	0.2809	−0.1134	−0.0018
"Snow"	0.1754	−0.0740	−0.0149
σ	0.4881	0.1027	0.0680
λ_j	0.0846	0.1123	0.8028



(a)

(b)

Fig. 6. Unified model's cut-off speeds (a) quantile, (b) Bayesian.

4.5 Example Illustration

The model that explains the variation in normalized speed using weather and visibility is shown in Eq. (9):

$$\begin{gathered}(\log(y)|\lambda_1,\lambda_2,\beta_1,\beta_2,\beta_3,\sigma_1,\sigma_2,\sigma_3)=\lambda_1\frac{1}{\sqrt{2\pi}\sigma_1}e^{\frac{\left(\log(y)-X^T\beta_1\right)^2}{2\sigma_1^2}}+\\ \lambda_2\frac{1}{\sqrt{2\pi}\sigma_2}e^{\frac{\left(\log(y)-X^T\beta_2\right)^2}{2\sigma_2^2}}+(1-\lambda_2-\lambda_1)\frac{1}{\sqrt{2\pi}\sigma_3}e^{\frac{\left(\log(y)-X^T\beta_3\right)^2}{2\sigma_3^2}},\end{gathered}\tag{9}$$

where vector X is the vector of weather conditions and visibility predictors, and y is the normalized speed. Here $(X^T\beta_1,\sigma_1)$, $(X^T\beta_2,\sigma_2)$, and $(X^T\beta_3,\sigma_3)$ are the locations and spreads of the mixture components and (λ_1,λ_2) are the mixture parameter.

Table 2 shows that the equations that govern the locations of the three components are

$$\begin{aligned}\mu_{\text{Congestion}} = -0.9025 + 0.0260 * \text{Visibility} - 0.0722 * \text{Medium Rain} - 0.0398 *\\ \text{Heavy Rain} + 0.2809 * \text{Freezing Rain} + 0.1754 * \text{Snow}\end{aligned}\tag{10}$$

$$\begin{aligned}\mu_{\text{Speed-at-capacity}} = -0.1947 + 0.0229 * \text{Visibility} - 0.0024 * \text{Medium Rain} -\\ 0.0465 * \text{Heavy Rain} - 0.1134 * \text{Freezing Rain} - 0.0740 * \text{Snow}\end{aligned}\tag{11}$$

$$\begin{aligned}\mu_{\text{Free-flow}} = 0.0335 + 0.0026 * \text{Visibility} - 0.0238 * \text{Medium Rain} - 0.0308 *\\ \text{Heavy Rain} - 0.0018 * \text{Freezing Rain} - 0.0149 * \text{Snow}\end{aligned}\tag{12}$$

We provide an example to show how to come up with the Q-quantile cut-off speed for a given weather group and visibility level. Based on the model, the predictors' vector is as shown in Eq. (13):

$$X^T = [\text{Visibility Medium Rain Heavy Rain Freezing Rain Snow}].\tag{13}$$

Assume the weather is "*FreezingRain*" and the visibility is "2"; what is the Q-quantile cut-off speed? Given the previous information, the predictors vector is shown in Eq. (14):

$$X^T = [12\quad 0\ 0\ 1\ \ 0].\tag{14}$$

Then the mean of speed-at-capacity component is calculated as shown in Eq. (15):

$$\begin{aligned}\mu_{\text{Speed-at-capacity}} = -0.1947 + 0.0229 * 2 - 0.0024 * 0 - 0.0465 * 0 - 0.1134 *\\ 1 - 0.0740 * 0.^{\cdot}\end{aligned}\tag{15}$$

Manipulating the above equation, we get −0.2623 as the mean of the speed-at-capacity component. Using the Matlab command “norminv(Q, −0.2623, 0.1123)” we get the Q-quantile cut-off speed where 0.1123 is the standard deviation for the speed-at-capacity component. Note that the standard deviation and the proportion parameters are constant and do not depend on the weather group or visibility.

Assume we are interested in the 0.001 quantile for “*FreezingRain*” and visibility is “2.” Using the Matlab command “norminv(0.001, −0.2623, 0.1123)” we get the 0.001 quantile cut-off speed, which is −0.6093. −0.6093 is the cut-off speed on the log scale, and thus the cut-off speed used to compute the binary matrix is exp(−0.6093) = 0.5437, which means the 0.001 quantile cut-off speed is 0.5437 of the posted speed. In other words, the cut-off speed is 0.5437 × 65 = 35.34 mph if the posted speed is 65 mph.

5 Conclusions

In this paper we propose an algorithm for real-time automatic congestion identification. The proposed algorithm models the speed distributions in free-flow, speed-at-capacity, and congested traffic states using a mixture of linear regressions. To the best of our knowledge, our proposed algorithm is the first methodology that considers the impact of weather and visibility in automated congestion identification. The parameters of the speed distributions were estimated using three different datasets covering three diverse regions, thus making this methodology more portable and transferable to any stretch of road in North America and potentially worldwide. The proposed algorithm is expected to be the state of practice and one of the routines used daily at many Departments of Transportation because of its simplicity, promising results, and suitability for running real-time scenarios. Because the proposed algorithm accurately identifies traffic congestion, both spatially and temporally, it is expected to be the state of practice pre-processing step toward identifying and ranking bottlenecks.

Acknowledgements. This effort was funded by the Federal Highway Administration and the Mid-Atlantic University Transportation Center (MAUTC).

Appendix A

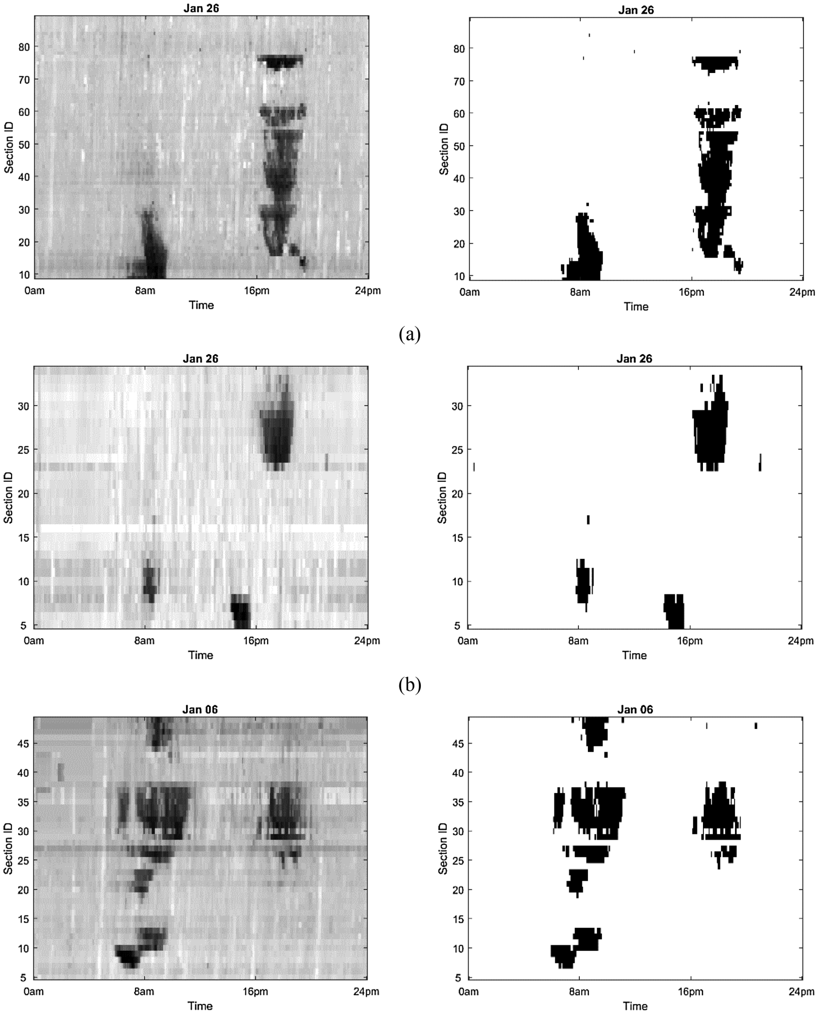
See Table 3

Table 3. Six weather groups.

Groups	#
Clear	1
Light Rain	2
Rain	3
Heavy Rain	4
Freezing Rain	5
Snow	6

Appendix B

Figure 7 shows the speed matrix and the corresponding binary matrix after applying the proposed algorithm. The binary matrix will be further filtered to fill gaps and remove noise using image-processing techniques.



(a)

(b)

(c)

Fig. 7. Speed (left) and binary matrix after applying algorithm (right); (a) Texas, (b) California, (c) Virginia.

References

1. Ibrahim, A.T., Hall, F.L.: Effect of adverse weather conditions on speed-flow-occupancy relationships. Transportation Research Record 1457 (1994)
2. Arnott, R., Small, K.: The economics of traffic congestion. Am. Sci. **82**(5), 446–455 (1994)
3. Sweet, M.: Does traffic congestion slow the economy? J. Plan. Lit. **26**(4), 391–404 (2011)
4. Jianming, H., et al.: Traffic congestion identification based on image processing. Intell. Transp. Syst. IET **6**(2), 153–160 (2012)
5. Guiyan, J., et al.: The method of traffic congestion identification and spatial and temporal dispersion range estimation. In: 2nd International Asia Conference on Informatics in Control, Automation and Robotics (2010)
6. Dailey, D.J., Trepanier, T.: The use of weather data to predict non-recurring traffic congestion. No. WA-RD 655.1. Washington Department of Transportation (2006)
7. Nookala, L.S.: Weather Impact on Traffic Conditions and Travel Time Prediction. University of Minnesota Duluth, Duluth (2006)
8. Chung, E., et al.: Does weather affect highway capacity. In: 5th International Symposium on Highway Capacity and Quality of Service. Yokohama, Japan (2006)
9. Brilon, W., Ponzlet, M.: Variability of speed-flow relationships on German autobahns. Transp. Res. Rec. J. Transp. Res. Board **1555**(1), 91–98 (1996)
10. Agarwal, M., Maze, T.H., Souleyrette, R.: Impacts of weather on urban freeway traffic flow characteristics and facility capacity. In: Proceedings of the 2005 mid-Continent Transportation Research Symposium (2005)
11. Hranac, R., Sterzin, E., Krechmer, D., Rakha, H.A., Farzaneh, M., Arafeh, M.: Empirical studies on traffic flow in inclement weather. Report No. FHWA-HOP-07-073 (2006)
12. Elhenawy, M., Rakha, H.A., Hao, C.: An automated statistically-principled bottleneck identification algorithm (ASBIA). In: 16th International IEEE Conference on Intelligent Transportation Systems (2013)
13. Sun, Z.-Q., et al.: Traffic congestion identification based on parallel SVM. In: Eighth International Conference on Natural Computation (2012)
14. Xu, L., Yue, Y., Li, Q.: Identifying urban traffic congestion pattern from historical floating car data. Procedia – Soc. Behav. Sci. **96**, 2084–2095 (2013)
15. Elhenawy, M., Rakha, H.: Automatic congestion identification with two-component mixture models. Transp. Res. Rec. J Transp. Res. Board **2489**, 11–19 (2015)
16. De Veaux, R.D.: Mixtures of linear regressions. Comput. Stat. Data Anal. **8**(3), 227–245 (1989)
17. Faria, S., Soromenho, G.: Fitting mixtures of linear regressions. J. Stat. Comput. Simul. **80** (2), 201–225 (2009)
18. Elhenawy, M., Chen, H., Rakha, H.A.: Traffic congestion identification considering weather and visibility conditions using mixture linear regression. In: Transportation Research Board 94th Annual Meeting (2015)
19. Saberi, M., Bertini, R.L.: Empirical analysis of the effects of rain on measured freeway traffic parameters. Transportation Research Board 89th Annual Meeting. No. 10-2331 (2010)
20. Smith, B.L., Byrne, K.G., Copperman, R.B., Hennessy, S.M., Goodall, N.J.: An investigation into the impact of rainfall on freeway traffic flow. In: 83rd Annual meeting of the Transportation Research Board, Washington DC (2004)